

WORKING P A P E R

A Description and Analysis of Evolving Data Resources on Small Business

AMELIA HAVILAND AND
BOGDAN SAVYCH

WR-293-ICJ

September 2005

This product is part of the RAND Institute for Civil Justice working paper series. RAND working papers are intended to share researchers' latest findings and to solicit informal peer review. They have been approved for circulation by the RAND Institute for Civil Justice but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.



Kauffman-RAND Center for the Study
of Small Business and Regulation

A RAND INSTITUTE FOR CIVIL JUSTICE CENTER

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 00 SEP 2005		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE A Description and Analysis of Evolving Data Resources on Small Business				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) RAND Corporation, 1776 Main Street, PO Box 2138, Santa Monica, CA 90407-2138				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 48	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

THE RAND INSTITUTE FOR JUSTICE

The RAND Institute for Justice is an independent research program within the RAND Corporation. The mission of the RAND Institute for Civil Justice, a division of the RAND Corporation, is to improve private and public decisionmaking on civil legal issues by supplying policymakers and the public with the results of objective, empirically based, analytic research. The ICJ facilitates change in the civil justice system by analyzing trends and outcomes, identifying and evaluating policy options, and bringing together representatives of different interests to debate alternative solutions to policy problems. The Institute builds on a long tradition of RAND research characterized by an interdisciplinary, empirical approach to public policy issues and rigorous standards of quality, objectivity, and independence.

ICJ research is supported by pooled grants from corporations, trade and professional associations, and individuals; by government grants and contracts; and by private foundations. The Institute disseminates its work widely to the legal, business, and research communities, and to the general public. In accordance with RAND policy, all Institute research products are subject to peer review before publication. ICJ publications do not necessarily reflect the opinions or policies of the research sponsors or of the ICJ Board of Overseers.

The Kauffman-RAND Center for the Study of Regulation and Small Business, which is housed within the RAND Institute for Civil Justice, is dedicated to assessing and improving legal and regulatory policymaking as it relates to small businesses and entrepreneurship in a wide range of settings, including corporate governance, employment law, consumer law, securities regulation and business ethics. The center's work is supported by supported by a grant from the Ewing Marion Kauffman Foundation.

A profile of the ICJ, summaries of its publications, and publications ordering information can be found on the Web at: www.rand.org/centers/icj

For more information on the RAND Institute for Civil Justice or the Kauffman-RAND Center for the Study of Regulation and Small Business, please contact:

Robert Reville, Director RAND Institute for Civil Justice RAND 1776 Main Street, P.O. Box 2138 Santa Monica, CA 90407-2138 (310) 393-0411 x6786; FAX: (310) 451-6979 Email: Robert_Reville@rand.org	Susan Gates, Director Kauffman-RAND Center for the Study of Regulation and Small Business RAND 1776 Main Street, P.O. Box 2138 Santa Monica, CA 90407-2138 (310) 393-0411 x7452; FAX: (310) 451-6979 Email: Susan_Gates@rand
--	---

PREFACE

This paper reviews the main existing and evolving data sources for research on small businesses and entrepreneurship in the U.S. While business data has traditionally focused on large firms, an increasing number of data sources include and identify small and new businesses. However, these data sources are scattered throughout government and private agencies. The primary aim of this review is to further educate researchers about these data sources in order to inform and promote research on small businesses and entrepreneurship. A secondary aim is to highlight gaps in available small and new business data to inform new or expanded data-gathering efforts and catalyze improved access to existing data sources. For each data source, the review summarizes the collection method, unit of observation, coverage, main variables, limitations, and current and potential future uses of the data. The review also includes a summary reference table concerning all the datasets described and references for accessing each data source and more detailed data-set specific information.

TABLE OF CONTENTS

Introduction	1
Government Data Sources: BLS	3
Quarterly Census of Employment and Wages (QCEW)	3
Business Employment Dynamics (BED)	5
Current Employment Survey (CES)	7
National Compensation Survey (NCS)	8
Current Population Survey (CPS)	10
Government Data Sources: Census Bureau	13
Standard Statistical Establishment Listing (SSEL) or Business Register (BR)	13
Longitudinal Business Database (LBD)	14
County Business Patterns (CBP)	15
Business Information Tracking System (BITS)	16
Economic Census (EC) and Company Organization Survey (COS)	18
Survey of Women/Minority Owned Business Enterprises (SWOBE/SMOBE), and Survey of Business Owners (SBO)	19
Integrated Longitudinal Business Database (ILBD) and Longitudinal Employer- Household Dynamic Program (LEHD)	20
Other Government Sources of Data	23
Survey of Small Business Finances	23
National Employer Health Insurance Survey	23
Medical Expenditures Panel Survey	24
Private and Commercially Available Data Sources	27
Duns Market Identifier (DMI)	27
Kauffman Firm Survey (KFS)	28
Research dataset derived from the Martindale-Hubbell Law directory	29
The Kaiser Family Foundation/Health Research and Educational Trust Employer Health Benefits Surveys	30
Discussion	30
Key Information Sources for Small Business Data	32
General Resources on Data for Small Businesses	32
Census Bureau Data Sources	32
Bureau of Labor Statistics Data Sources	33
Other Government Data Sources	33
Private Sources of Data	34
Conference Resources:	34
NAS Panels:	34
References	35
Appendix A	38
Description of Main Parts of the Table	38

ACKNOWLEDGEMENTS

We thank Susan Gates and Michael Greenberg for their insightful suggestions.

INTRODUCTION

Historically, data collected on U.S. businesses have focused almost exclusively on large firms (typically those with at least 250 employees). As a result, researchers interested in small businesses and entrepreneurship have been strongly constrained in their ability to carry out empirical, policy-related research. Ongoing concerns about the lack and quality of data on small firms led to a recent conference on data sources related to entrepreneurship¹ and to the creation of a National Academy of Sciences panel on Federal Business Statistics, which will issue its final report in 2006. Such efforts have been part of a more recent trend to increase the number, quality, and richness of data sources on small firms. But while data sources continue to improve, information on the uses, availability, and limitations of these sources is scattered among the multitude of governmental and private organizations that collect and own the data.

To study the position and role of small businesses, researchers need access to various kinds of datasets, including cross-sectional and longitudinal data, as well as data collected at the level of firm, establishment, owner, or worker. If a necessary dataset does not yet exist, researchers may instead need to carry out their own data collection using a sampling frame of firms or establishments. Ideally, this sampling frame would provide enough information to identify units of interest without screening. At present, there continue to be significant challenges involved in obtaining longitudinal data, data that include information at several levels, and sampling frames. Two of the main problems are lack of availability due to cost or confidentiality concerns, and the poor quality of linkages, either within units over time or between establishments and firms. Other important concerns for longitudinal data sources include the point in time at which a new firm or establishment is identified to enter a database, the point in time at which it is determined that a business has closed and should be removed, and the point in time at which the size of the business, whether based on number of employees or some other criteria, is measured.²

This document briefly describes the main government and private data sources currently available or under construction for research on small business and entrepreneurship. The paper also provides a listing of resources researchers can use to gain more information about each dataset. Of special note, information is provided on two relatively new government longitudinal databases, the Longitudinal Business Database (LBD) and the Business Information Tracking System (BITS), both of which include establishment and firm linkages. The paper also describes the Kauffman Firm Survey, a private survey going into the field in 2005, which will provide publicly available longitudinal data on new firms. In

¹ Kauffman Symposium on Entrepreneurship Data (Nov 10-11, 2004).

² See the following reference for information on how size class measurement timing can matter, "Why size class methodology matters in analyses of net and gross job flows." Coredeia Okolie; Monthly Labor Review, July 2004. <http://www.bls.gov/opub/mlr/2004/07/art1abs.htm>

addition, the document summarizes the types of data currently available, issues researchers need to be aware of, and data needs that are not yet being met.

For each dataset we describe the collection method, coverage, main variables and limitations, and data uses. We start the description of each dataset with a discussion of how the data are obtained, whether from administrative records, survey, or census. Then, we identify the population from which the data are collected (e.g., firm, establishment, owner, or worker) and unit of observation. Most of the business information is collected at the level of the establishment; a single location where goods are produced or services are provided. In some cases, these establishment data can be aggregated to the level of firm, which might include several establishments. We also describe the main variables available in the dataset, and the main limitations of the data. We are especially interested in the variables that measure size of the firm because of the obvious relevance to small business research. We select employment as the main measure of the firm size, although other measures like revenues or sales are possible and are available in some datasets. We conclude the description of each dataset with a discussion of current and potential future uses of the data. Table 1 of the appendix provides a summary reference concerning the datasets described in this report.

This document is organized into four sections. The first three sections discuss government data sources: the Bureau of Labor Statistics, the Census Bureau, and other government sources. The fourth section describes private data sources. These sections are followed by a brief discussion of ongoing data collection needs, and a lightly annotated bibliography of further references on the data sources described here and recent research on small businesses that make use of these data sources.

GOVERNMENT DATA SOURCES: BLS

A multitude of government agencies collect information relevant to small business research. The following section focuses on several datasets created by the Bureau of Labor Statistics. The Bureau of Labor Statistics (BLS) is the principal Federal agency that collects data in the field of labor economics and statistics. The mission of the BLS is to “collect, process, analyze and disseminate essential statistical data to the American public, the U.S. Congress, other Federal agencies, State and local governments, business, and labor.”³ Its goal is to provide timely, consistent and high-quality data on a range of issues including employment, wages and workers’ benefits. To achieve its mission, the BLS collects a number of different datasets. Those include the Quarterly Census of Employment and Wages, Business Employment Dynamics, Current Employment Statistics, and National Compensation Survey. The BLS also cooperates with the Census Bureau to collect data for the Current Population Survey. Although each of the surveys have somewhat different coverage, collection procedure, and available variables, all of these datasets can be useful in identifying the members of the small businesses universe.

QUARTERLY CENSUS OF EMPLOYMENT AND WAGES (QCEW)

Data Collection Method and Coverage

The Quarterly Census of Employment and Wages (QCEW), also known as the Covered Employment and Wages (CEW) and ES-202 program, is the most extensive effort by BLS to collect quarterly employment and payroll information. This comprehensive dataset is the result of cooperation between the BLS and the State Employment Security Agencies (SESAs). The State Employment Security Agencies derive data from quarterly contribution reports submitted by employers. All employers are required to pay quarterly taxes based on the employment and wages for workers covered by state unemployment insurance (UI) laws and federal workers covered by the Unemployment Compensation for Federal Employees (UCFE) program. The data are derived from the administrative records that employers need to submit to the SESAs for each of their establishments.

As such, the data include all establishments subject to state and federal unemployment insurance laws. In 2004, the data were collected from about 8.4 million establishments. As described in the BLS’s handbook of methods, QCEW provides a virtual census of nonagricultural employees and their wages, and, additionally, about 47% of all workers in agricultural industries are covered (see BLS, 1997). The data cover approximately 98 percent of all employment;⁴ the major exclusions from UI coverage are self-employed workers, religious organizations, most agricultural workers on small farms, active duty military

³ See the BLS mission statement at <http://www.bls.gov/bls/blsmisn.htm>, accessed (July 5, 2005).

⁴ See technical note to the most recent report <http://www.bls.gov/news.release/cewqtr.toc.htm> (Accessed June 15, 2005).

personnel, elected officials in most states, most employees of railroads, some domestic workers, unpaid family workers, most student workers at school and certain employees of certain small nonprofit organizations (note that the regulations determining which nonprofits are required to have UI coverage vary by state and may change over time).

Main Variables

The QCEW dataset includes information on establishment's monthly employment, quarterly wages, location (full address), and industry classification. Monthly employment data represent the number of workers who worked during, or received pay for, the pay period including the 12th day of each month. As explained in BLS (1997), the employment measure includes all corporation officials, executives, supervisory personnel, clerical workers, wage earners, pieceworkers, and part-time workers. Workers are reported in the state and county of the physical location of their job. This measure of employment includes all workers on paid sick leave, paid holiday, paid vacation, and so forth, but excludes those on leave without pay for the entire payroll period (see BLS, 1997).

The QCEW also provides data on quarterly wages; however, the definitions of what constitutes wages differ between states. In most states, total wages include gross wages and salaries, bonuses, stock options, tips and other gratuities, and the value of meals and lodging, where supplied. Some states also include in total wages employer contributions to certain deferred compensation plans, such as 401(k) plans.

All variables in the QCEW are collected and provided at the level of establishments. An establishment is an economic unit that produces goods or provides services, such as a factory, mine or store. Usually, it is a single physical location engaged in predominantly one economic activity (see BLS, 1997). Employers who operate multiple establishments within a state submit a Multiple Worksite Report that provides detailed information on each of the establishments. Although the data are provided at the level of establishment, one can aggregate it to the level of firm using Employer Identification Numbers assigned by the IRS. See more details in the discussion of the Business Employment Dynamics (BED) database.

Major Limitations

The primary disadvantage of this dataset is lack of accessibility. The dataset, as well as number of other datasets constructed by BLS, is not publicly available, but can be accessed only in restricted settings at the BLS offices and after an application process.⁵ BLS accepts applications for use of the dataset four times a year, and any research must benefit the BLS. Both the application and the application time are relatively short, with a 5-10 page proposal due on April 15th for use of the data starting in the summer. In

⁵ For the list of available datasets and application rules see <http://www.bls.gov/bls/blsresda.htm> (accessed June 8, 2005). The BLS refers to the QCEW and the BED as the Longitudinal Database of Establishments Covered by State Unemployment Insurance Programs.

addition, there are restrictions on the type of analysis outputs from the QCEW that can be released. Although the dataset is not publicly accessible, extensive aggregate tables created from these sources are publicly available. The QCEW program provides tables for employment and wages aggregated by geography (Nation, State, MSA, County) and industry.⁶ For the first quarter of each year, BLS also produces these tabulations by establishment employment categories. Researchers using the publicly available aggregated tables should be aware that some information is withheld by the BLS to prevent disclosure of individual employers.

In addition, researchers using the dataset need to be aware of differences between states in who is covered and what data are collected. Also, the year-to-year consistency of the employment and wage data might be affected by periodic changes in state and federal UI laws. For example, since January 1, 2004, the Washington Employment Security Department no longer includes as covered wages an employee's income attributable to the transfer of shares of stock to the employee. The details of coverage and differences between states are provided in the description of the state unemployment insurance legislation.

Furthermore, the QCEW dataset is not designed as a time series. However, the QCEW establishment level observations can be connected over time to construct a longitudinal database. These linkages were used to create the Business Employment Dynamics database described below.

What Can be Done with the Data

The QCEW provides cross-sections of administrative data for a limited number of variables covering nearly the entire the universe of establishments. Records collected through the QCEW program can be used in a number of situations. The BLS links these cross-sectional records to produce a longitudinal database of establishments (see discussion of the Business Employment Dynamics program below). The Census Bureau also draws upon these data to supplement various multilevel business databases. In addition, this database can be used as a sampling frame for further surveys, and to produce denominators for other research, if access issues are resolved.

BUSINESS EMPLOYMENT DYNAMICS (BED)

Data Collection Method and Coverage

The Businesses Employment Dynamics (BED) program connects establishment records from the QCEW administrative records over time to produce the Longitudinal Database of Establishments Covered by State Unemployment Insurance Programs, also known as Business Employment Dynamics (BED) dataset.⁷ The data are complete from September 1992-2003 and are being expanded on an ongoing basis.

⁶ See <http://www.bls.gov/cew/> (accessed June 8, 2005).

⁷ Some researchers also refer to this database as BLS Longitudinal Database or LDB (see Pivetz, Searson, and Slpetzer, 2001).

These longitudinal data are more frequently updated and available on a timelier basis than anything previously available.

Pivetz, Searson, and Spletzer, (2001) describe how the BED longitudinal linkages are developed. Most of the establishments are matched over time using the unique SESA identification number (SESA-ID). This approach, however, might miss possible links because of changes in ownership, firm restructuring, or UI account restructuring. In this case, probability based matching is used to link establishments with different SESA-IDs. The match is based upon comparisons such as the same name, address, and phone number. Third, an analyst examines unmatched records individually and makes a possible match. The resulting dataset includes longitudinal histories of over 6.4 million private sector employer reports.

Given that the dataset is derived from the QCEW establishment records, one should not expect large differences in coverage and definition of variables between QCEW and BED. The major contrast to the QCEW data is that the BED dataset excludes government employees, private households, and establishments with zero employment.⁸

Main Variables

Most of the variables in the BED are the same as in the QCEW: monthly establishment employment, quarterly wages, mailing and physical address, and industry classification. Using these data, one can track net employment changes at the establishment level over time, and determine job gains and losses at expanding and contracting establishments.

More recent work has also presented BLS establishment level data aggregated to the firm level using Employer Identification Numbers assigned by the IRS (See Okolie, 2004). Using the BED data, Okolie (2004) presented net employment changes at both the establishment and firm level. The underlying BED data are presented at the establishment or the UI reporting unit level. Aggregating to the level of firm using Employer Identification Numbers helps link firms that are operating in multiple states. Although these firms would have separate UI accounts for each state, they will have one employer identification number covering all of the establishments across the country. The EIN, however, is an imperfect way to aggregate to the firm level. A multiunit firm can be associated with a cluster of one or more EINs. This distinction is not clear in the BLS data.

Major Limitations

A major limitation of this database is that it is not publicly available. Interested researchers need to apply for access to this database through the BLS research center.⁹

⁸ See technical notes to the most recent BED press release <http://www.bls.gov/news.release/cewbd.toc.htm> (accessed June 8, 2005)

⁹ For the list of available datasets and application rules see <http://www.bls.gov/bls/blsresda.htm> (accessed June 8, 2005).

Another limitation of this dataset is that a significant change in the reporting does not allow for direct comparisons of the records before and after 1991. Prior to 1991, employers provided data on a reporting unit basis, where each reporting unit provided information for different county locations within a state. However, these units did not necessarily coincide with establishments.

While these data allow for identification of openings, closings, expansions, and contractions of establishments, there is likely to be a lag time in correctly identifying closings due to state agencies' tendency to impute employment figures for as many as two quarters after zero employment is reported.

What Can be Done with the Data

Data from the BED program can help provide a dynamic picture of local labor markets. It is an important new database that includes the universe of longitudinal establishment data. As such, these data make it possible to study within-establishment changes over time and across location. For example, Faberman (2002) used these data to study different aspects of job creation and job destruction. In addition, these data can be used to answer basic questions about changes in establishment and firm sizes over time. For instance, it is possible to study growth in firms and establishments by employment size categories recorded at different points in the year. This information is currently unavailable and could change inferences based solely on March employment or average annual employment.

CURRENT EMPLOYMENT SURVEY (CES)

Data Collection Method and Coverage

The BLS conducts several surveys that provide monthly information on labor market conditions, including the current employment statistics program (CES). As part of the CES, the BLS cooperates with the State Employment Security Agencies (SESAs) to collect monthly establishment level data on employment, hours and earnings. SESAs obtain this information from a sample of about 160,000 firms and government agencies monthly (about 400,000 establishments). All firms with 1,000 or more employees are asked to participate in the survey, as is a sample of firms across all employment sizes. The CES survey draws most of its sample from the UI administrative records. The sample excludes all agriculture, private households, and self-employed workers. In addition, it includes some jobs not covered in QCEW. For example, it includes employees of railroads and religious organizations, as well as some other non-UI-covered jobs. As of 2003, the sample design is a stratified, simple random sample of establishments from the BLS's Longitudinal Data Base, with strata defined by state, industry and employment size.

Main Variables

The survey obtains information from each establishment on the total number of employees, number of women employees, and number of production workers, their payroll and hours worked. It also

gathers information on several kinds of changes in employment or wages as well as reasons for these changes. For employment measures, data are collected on seasonal changes, short-term projects, layoffs, strikes, temporary shutdowns, or internal reorganization. For wage changes, data are collected on shifts in the wage rate, change in the quality of personnel, changes in the hourly pay or incentive pay, and overtime. In addition, data are collected on location and industry specialization.

Major Limitations

This survey has several limitations. First and foremost, the survey is not publicly available, but is available only for use by approved researchers at the BLS. From these data, a large number of employment, hours, and earnings series in considerable industry and geographic detail are prepared and published each month, although none are available by the categories of establishment or firm size.

Another issue is that the dataset is not longitudinal, although the sample design uses rotating samples of establishments that follow some firms for about 4 years. The possibility of linking these establishments over time has not yet been discussed. In addition, survey non-response may be a problem. For example, Copeland (2003) notes that depending on the industry, complete responses are obtained for only about 47%-57% of the sample.

In addition, there was a major redesign to the CES survey in 2003, which might make comparisons over time somewhat problematic. In particular, there was change in the sample design. It was changed from an historical quota sample design to a probability sample.

What Can be Done with the Data

To date, little work on the state of small businesses has been done with these data. Researchers might prefer using information derived from the QCEW program that is based on full administrative records instead of a survey sample. However, the CES provides information that is not available in QCEW and that could be important for tracking economic development, such as reasons for changes in numbers of employees, and the number of women and production workers.

NATIONAL COMPENSATION SURVEY (NCS)

Data Collection Method and Coverage

The BLS conducts a number of surveys aimed at obtaining detailed information about compensation and benefits. Recently, a number of compensation surveys were combined into the National Compensation Survey program (NCS).¹⁰ Since 2000, the NCS survey has collected information that was previously collected using the Occupational Compensation Survey (OCS), the Employment Cost

¹⁰ For a description and further information about the NCS program see the program's website <http://www.bls.gov/ncs/>, BLS Handbook of Methods (BLS, 1997), Cohen (1997), and Bostin (2004).

Index (ECI), and the Employee Benefit Survey (EBS). The NCS provides detailed measures of occupational earnings, compensation cost trends, benefits, and detailed plan provisions.

The NCS is fielded every year and includes about 18,000 establishments, which are selected in a three-stage design. In the first stage, regions of interest are selected. In the second stage, establishments within these areas are chosen, with the sample frame stratified by ownership and industry. The list of establishments is derived from the QCEW records. Before 1999, NCS included only establishments with 50 or more workers. Beginning in 1999, the survey has covered establishments with one or more workers and state and local public agencies with 50 or more workers. Federal government, agriculture, and private households are excluded from coverage. According to the technical note to the NCS report, each sampled establishment was selected within a stratum with a probability proportional to its employment (BLS, 1997). In the third stage, a probability sample is taken of occupations within an establishment. Compensation for workers of specific occupations is collected.¹¹

Main Variables

The data include variables on establishment employment, full address, industry classification, establishment characteristics, workers' variables including occupational details, full-time and part-time status, hours worked, wages, union status, and type of pay arrangements. The survey also collects information on various benefits, such as health, life, and disability insurance, medical premiums paid, retirement, leave information, and overtime pay.

Major Limitations

There are several major limitations to using this survey for the analysis of the small business affairs. One limitation is that the data are not publicly available. Researchers need to apply for access to this database through the BLS research center.¹² Some aggregate tables for localities, broad regions, and the nation by establishment employment size category are available.¹³

In addition, the sample is selected in such a way that larger firms are more likely to be surveyed. Furthermore, there were changes over time in the data collection methods. Prior to 1999, establishments of different size were targeted in different years – data for medium and large establishments (more than 100 employees) were collected in odd years, and for small establishments in even years.

What Can be Done with the Data

NCS data have been used to examine factors that determine low-wage labor (Bernstein and Gittleman, 2003); incidents of provision of health benefits (Barsky, 2004); and trends in employer-

¹¹ See <http://www.bls.gov/ncs/methodology.htm> (last accessed July 5, 2005).

¹² For the list of available datasets and application rules see <http://www.bls.gov/bls/blsresda.htm> (accessed June 8, 2005).

¹³ See <http://www.bls.gov/ncs/> (accessed July 5, 2005).

provided prescription-drug coverage (Dietz, 2004). Each of these topics could be addressed by establishment employment size categories. With these data, it also would be possible to evaluate the relationship between employer-provided health benefits and establishment or worker characteristics by establishment size.

CURRENT POPULATION SURVEY (CPS)

Data Collection Method and Coverage

In addition to establishment-level data, one of the BLS' most widely used household surveys, the Current Population Survey (CPS), also can be used to identify some members of the small business universe. The United States Census Bureau conducts the Current Population Survey (CPS) for BLS. The microdata from this survey are publicly available with coverage beginning in 1962. The BLS uses the data to provide monthly estimates of the number of unemployed people in the United States. The CPS also serves as a vehicle for supplemental studies on subjects other than employment. In contrast to other surveys we have discussed, the CPS uses the household, rather than the establishment or firm, as its sampling unit. The survey nonetheless provides a comprehensive current source of information on the occupation of workers and the industries in which they work. However, the survey provides only a limited amount of firm-level information. For these reasons, the CPS can be a particularly good source of information about the self-employed and new entrepreneurs.

The CPS collects information on the labor force status of the civilian non-institutional population 15 years of age and older. The data are obtained from a sample of about 60,000 households. The CPS stands out as having consistently very high response rates among government surveys. The households are selected by a multistage stratified statistical sampling scheme.¹⁴ The data for all members of the household are recorded in separate records. The sample is selected to assess overall employment, unemployment, and the number of people in and out of the labor force. This sample includes categories of workers that are entirely or partly excluded from the QCEW program. For instance, people are classified as employed if they did any work at all as paid employees during the reference week; worked in their own firm, profession, or on their own farm; or worked without pay at least 15 hours in a family firm or farm. People are also counted as employed if they were temporarily absent from their jobs because of illness, bad weather, vacation, labor-management disputes, or personal reasons.

Researchers can take advantage of the panel structure of the CPS. Every housing unit in the CPS is interviewed for four consecutive months and then dropped out of the sample for the next eight months and brought back in the following four months. In most years the observations can be linked over time using household identification number. However, these links can generate false positives because of non-

¹⁴ Details about CPS methodology are provided in BLS (2002)

response, migration, mortality or recording errors. Algorithms are available to improve matching using changes in gender, race, age, and/or educational attainment. Tradeoffs of these algorithms are discussed in Madrian and Lefgren (1999).

Main Variables

For each member of the household, the basic CPS monthly survey collects demographic information, such as age, sex, race, marital status, veteran status, Hispanic origin, immigrant status, educational attainment, and family structure. In addition, information is collected on labor market outcomes: employed, unemployed, searching or not searching for work. The reference period for labor market information is defined as the 7-day period (from Sunday through Saturday) that includes the 12th of the month. Respondents are also asked questions about class of worker (private, government, self-employed, without pay, and never worked), hours worked in reference week, occupation, industry type, reasons for working part-time, and reasons for lack of employment. Respondents in the outgoing rotation panel (those households that are in the panel for month number 4 or number 8) are asked about earnings in their main job. The self-employed are usually asked a similar set of questions as the other workers. New business owners can be identified by matching two of the multiple time points at which an individual is included in the survey, either across months or across years, and flagging those who are business owners at the second but not the first time point.

In addition to the demographic and labor force information mentioned above, the March demographic supplement survey collects information on firm employment,¹⁵ health insurance, and other non-cash benefits provided.

Other supplements on selected topics are included for most months. These supplemental topics are often repeated in the same month from year to year. For example, biennial September supplements gather detailed information for veterans, their service-connected income, effects of a service-connected disability on current labor force participation, and participation in veterans' programs. Biennial February supplements collect information about type of employment arrangement workers have on their current job and other characteristics of the current job such as earnings, benefits, longevity, employee satisfaction rates and expectations. Additional supplements collect information about computer and Internet use at home and at work, job tenure and occupation mobility, adult education, and health and pension coverage.

Major Limitations

The CPS has several limitations, the most important of which for our purposes is that it contains limited information about firms. The CPS is a household survey designed to collect unemployment data, and very little information is collected about firms where respondents work or which they own. Other

¹⁵ The specific wording of the question in the March supplement is: NOEMP - Counting all locations where this employer operates, what is the total number of persons who work for '[the respondent's]' employer? Response categories are: Under 10, 10–24, 25–99, 100–499, 500–999, and 1000+.

possible disadvantages of the CPS are that it covers only a two-year longitudinal panel, loses a significant portion of the sample due to matching algorithms, or moving of households, and makes comparability over time a problem because of significant redesigns. In addition, major redesign of the survey instrument in January 1994 might constrain comparison of some series before and after 1994.

What Can be Done with the Data

The overall advantages of the CPS are large sample sizes, long time series, quick access to timely data, a very large built-in comparison group of non-entrepreneurs, and a wide range of topics in the supplements. The coverage of the CPS makes it an important tool for the analysis of the self-employed and new entrepreneurs. There are a number of papers that use the CPS for the analysis of issues that are relevant for small businesses. For example, Berger et al. (1999) used March CPS data to examine the distribution of low-wage workers by firm employment size categories, as well as effects of the minimum wage. Labor economists have used the CPS extensively for the analysis of the relations between employment size of a firm and wages (see Idson and Feaster, 1990; Mellow, 1982; Bowlus, Kiefer and Neumann, 1995; Pearce, 1990; Card, 1996; Brown and Medoff, 1989; Antos, 1983; Hirsch and Schumacher, 1998; Weiss and Landau, 1984; Evans and Leighton, 1989). Other studies include analysis of the prevalence of formal on-the-job training (Loewenstein and Spletzer, 1997); factors that explain differences in turnover between large and small firms (Even and Macpherson, 1996); effects of health insurance on hours worked (Cutler and Madrian, 1998; Gruber and Poterba, 1994); effects of employment protection (Oyer and Schaefer, 2002); earnings by racial or ethnical characteristics (Agesa, Agesa, and Hoover, 2001; Carrington, McCue, and Pierce, 2000; Trejo, 1997); patterns of entrepreneurship (Evans and Leighton, 1989); gender differences in earnings (Sorensen, 1990; Macpherson and Hirsch, 1995); evidence of labor market cycles for the self-employed (Carrington, McCue, and Pierce, 1996); transition between full-time, part-time and retirement (Peracchi and Welch, 1994); worker compensation (Hirsch, Macpherson, and DuMond, 1997); patterns of self-employment among older U.S. workers (Karoly and Zissimopoulos, 2003); and access to computers and the decision to become self-employed (Fairlie, 2005).

The CPS can be used to support further studies of self-employment and entrepreneurship, including patterns of health insurance coverage, human capital, and education among the self-employed. In addition, the CPS can be used to examine patterns of self-employment and entrepreneurship among recent immigrants or other demographic groups of interest.

GOVERNMENT DATA SOURCES: CENSUS BUREAU

The Census Bureau is the leading source of quality data about the nation's people and economy. It provides an extensive collection of data on businesses and people. The Constitution and Congress mandate some of the data collection. For example, the Census Bureau has a mandate to conduct population censuses every ten years and economic censuses every five years. Business data at the Census Bureau include the Business Register, Business Information Tracking system, Integrated Longitudinal Business Database, Economic Census, Company Organization Survey, Survey of Women/Minority Owned Business Enterprises and Survey of Business Owners. Although these surveys are available only through the Census Research Data Centers, they provide information on different dimensions of establishments and firms. Importantly, researchers can combine many of these sources to provide more detailed information about the small business universe.

STANDARD STATISTICAL ESTABLISHMENT LISTING (SSEL) OR BUSINESS REGISTER (BR)

Data Collection Method and Coverage

The Standard Statistical Establishment Listing (SSEL), also known as Business Register (BR), contains records for each known establishment and company that is located in the United States. The detailed cross-sectional data are available for each year since 1975, although the frequency of updating the data varies depending on the industry. The BR systematically incorporates information about firms and establishments from a number of administrative sources, censuses, and surveys.¹⁶ The first source is payroll tax information provided by the IRS, which provides employment and payroll information. The second source is for new firms and is provided by the Social Security Administration. The data come from applications for Employer ID numbers (EINs) filled out by all new employers. The Business Establishment List, separately created and maintained by the BLS from unemployment insurance administrative records, is used to fill in industry classifications (SIC codes) for establishments otherwise missing this information. The information is regularly updated using the Company Organization Survey, an annual survey of manufacturers, and the economic census.

As described on the Census Bureau Web Site,¹⁷ this dataset covers establishments of all domestic employer and non-employer businesses (except private households and governments) and establishments that are parts of multi-establishment firms. The cross-section includes 180,000 multi-unit companies, representing 1.5 million affiliated establishments, 5 million single-establishment companies, and nearly 14 million non-employer businesses.

¹⁶ Description of the BR is available at <http://www.census.gov/econ/overview/mu0600.html> (accessed June 20, 2005).

¹⁷ See <http://www.census.gov/econ/overview/mu0600.html> (accessed June 20, 2005).

This data source is used to create the Statistics of U.S. Businesses (SUSB) and the County Business Patterns (CBP), both of which are useful sources because the aggregate statistics they contain are made publicly available at fine geographic and industry levels and by establishment employment size.

Main Variables

For each establishment record, the BR has information on the number of employees, receipts, full address, company affiliation information, and industry classification. The employment and payroll data are continuously updated using various administrative sources. The data also include identifiers that can be used to connect establishments to firms and units over time. Those are Census File Number (CFN), Employer Identification Number (EIN), and permanent plant number (PPN). The Business Register provides all the information for each establishment as well as for the firm.

Major Limitations

One of the major limitations of this dataset is that it is not publicly available. Researchers can apply for access to the microdata at the Census Research Data Centers.¹⁸ In addition, aggregated tables from the data are available through County Business Patterns. Additional tables also can be requested and are provided at a cost relative to the work involved in creating the table.

What Can be Done with the Data

The BR provides underlying records that form a number of statistical databases within the Census Bureau. For example, the Longitudinal Business Database and Business Information Tracking System connect some of the records from the BR over time to create longitudinal databases of establishments. In addition, the BR serves as a sampling frame for a number of surveys conducted by the federal government.

LONGITUDINAL BUSINESS DATABASE (LBD)

Data Collection Method and Coverage

The Center for Economic Studies of the Census Bureau links the annual snapshot files from the BR over time to create the Longitudinal Business Database (LBD). Jarmin and Miranda (2002) describe in detail how the LBD is constructed. In particular, the matching algorithm relies on available numerical identifiers as well as information about name and address of the firm to make linkages over time. Three numerical identifiers are available in the LBD. Census File Number (CFN) is used by the Census Bureau to identify establishments in economic censuses and surveys. However, this number can change over time due to changes in ownership status, in single and multi-unit status, or in the legal form of

¹⁸ For detailed description see <http://www.ces.census.gov/ces.php/rdc> (accessed June 20, 2005).

organization. The Census Bureau tracks changes in the CFN. Another identifier is the Permanent Plant Number (PPN). This number was introduced to BR in 1982 and is designed to stay unchanged as long as the establishment remains active at the same location. The other available identifier is the Employer Identification Number (EIN). This is a taxpayer identification assigned by the IRS. As described in Jarmin and Miranda (2002), longitudinal matching first uses information in PPN, but if there were no matches the matching was done using CFN, and, after that, EIN and possibly tracking changes to CFN. Matching is also augmented using the name and address of the firm.

The resulting longitudinal database covers nearly all of the non-farm private economy, as well as some public sector activities from 1975 to the present. It includes about 4.5-7.1 million records per year, for a total of almost 24 million unique establishments from 1975 to the present. It excludes establishments with zero annual payrolls.

Main Variables

The LBD draws its main variables from the BR. Those include establishment and firm employment, payroll, revenues, full address, firm affiliation, and industry classification. In addition, the LBD includes information on the enterprise age and tenure. There are also plans to include variables on ownership status and ownership changes.

Major Limitations

Like other microdata from the Census Bureau, this dataset is not publicly available. Researchers can apply for access to the microdata at Census Research Data Centers.¹⁹ In addition, the accuracy of the firm to establishment links varies over time: quality of the links declines after an Economic Census and then improves again with the next census (Jarmin and Miranda, 2002).

The LBD is useful to researchers who want to examine entry, exit and gross job flows by establishment or firm employment size. The data allows researchers to study changes over a long time frame within establishments. In the past, the dataset was used to examine entry and exit of firms in specific industries (Jarmin et al., 2004) and establishment and employment dynamics (Foster, 2003). This dataset can be connected to other Census Bureau products.

COUNTY BUSINESS PATTERNS (CBP)

Data Collection Method and Coverage

County Business Patterns (CBP) is one of the programs that provide geographic aggregates of micro-level establishment data. The CBP aggregates data from the BR to the level of county according to

¹⁹ For detailed description see <http://www.ces.census.gov/ces.php/rdc> (accessed June 20, 2005).

industry and establishment employment size category.²⁰ However, it does not provide data by firm size categories. Yearly tables are provided for data from 1964 to the present.

The CBP covers all establishments with paid employees included in the BR, although it excludes some industries. In particular, the CBP excludes crop and animal production; rail transportation; National Postal Service; pension, health, welfare, and vacation funds; trusts, estates, and agency accounts; private households; and public administration. The CBP also excludes most government establishments.

Main Variables

The CBP provides aggregate tables for employment during the week of March 12, total number of establishments, first quarter and annual payroll by industry, and establishment employment size class. The data on total employment and payroll are suppressed whenever they would disclose the operations of an individual employer. In addition, the Zip Code Business Patterns Data provide the number of establishments by industry codes and establishment employment categories for each zip code.

Major Limitations

Although this product is publicly available, it provides only aggregated tables. These might be less useful in many instances in which firm- or establishment-level data are necessary. In addition, the CBP does not provide information by firm employment size, only by the establishment size.

What Can be Done with the Data

The CBP dataset is a standard reference source of local economic data. The CBP tables are often used to derive denominators for employment and number of establishments in a particular establishment size category (numerator data typically come from other sources).

BUSINESS INFORMATION TRACKING SYSTEM (BITS)

Data Collection Method and Coverage

Over the years there have been several efforts to link records in the Business Register over time to create a longitudinal database of establishments.²¹ The Business Information Tracking system (BITS), also known as the Longitudinal Enterprise and Establishment Microdata (LEEM), is one of the most recent efforts to create a longitudinal database. The Census Bureau has collaborated with the Small Business Administration (SBA) Office of Advocacy to create the Business Information Tracking System (BITS), which links annual County Business Patterns data over time from 1989 to 2001 (and ongoing with a two-year lag) and within firms. As described in Acs and Armington (2005), the primary links over time are constructed using the Census File Number (CFN) – the Census Bureau Identification number for

²⁰ See description at <http://www.census.gov/econ/overview/mu0800.html> (accessed June 20, 2005).

²¹ Also see description of the Longitudinal Business Database.

establishments in the Economic Census. However, this number would not connect establishments that changed ownership or organizational structure between years. These establishments are matched using a permanent plant number (PPN) – an establishment identifier that does not change over time, Employer Identification Number (EIN), and other establishment attributes like name, address, zip code and industry codes. Robb (1999) and Acs and Armington (1998) provide documentation for the BITS data.

There are several key differences between the LBD and the BITS. First, the BITS has more restrictive coverage of industries. It uses the same coverage as the CBP. In particular, the coverage excludes some agricultural industries, railroads, postal service, private households, large pension, health, and welfare funds and public administration. In addition, the LBD have longer panels. While the BITS goes back only until 1988, the LBD links establishments back to 1976. On the other hand, BITS currently includes both firm and establishment data, while both levels of data are not yet available for the LBD.

Main Variables

The BITS includes establishment-level information on employment (missing for 15-18 percent), annual payroll, location information including state, MSA, city, and place codes; start year, and industry code. The records also include firm-level information on aggregate employment, payroll, primary industry, and primary location (based on the establishment with the largest number of employees).

Major Limitations

The BITS data are not publicly released, but researchers can apply to use the data at Census Research Data Centers (RDCs). Researchers are encouraged to work with staff at one of the RDCs to develop a research proposal, which is likely to take a minimum of six months and sometimes substantially longer to be reviewed. The proposed research must further the interests and goals of the Census Bureau.²²

The largest changes in the BITS data occur in and just before Economic Census years when industry classifications are clarified and changes in establishments within firms are recorded.

What Can be Done with the Data

Using this longitudinal database, it is possible to identify establishment births, deaths, expansions, and contractions. Most census products can be connected to each other using an EIN or PPN. Therefore, information from the other Census products could be used to identify a type or category of establishments, and BITS can be used to track those establishments over time. There are several studies that used BITS data to examine issues that may be important for small businesses. The data were used to examine the persistence of new jobs (Acs and Armington, 2004); job flow dynamics (Acs and Armington,

²² For more information see <http://148.129.75.149/ces.php/guidelines>

2000; Armington, Robb, and Acs, 1999); survival of firms in various industries, including start-ups (Headd, 2001; Boden, 2000); and mergers and acquisitions (White, 2002; Armington and Robb, 1998).

ECONOMIC CENSUS (EC) AND COMPANY ORGANIZATION SURVEY (COS)

Data Collection Method and Coverage

Two Census products central to the Business Register, the BITS and the LBD, are the Company Organization Survey (COS), also known as Report of Organization Survey, which is fielded annually except in Economic Census years, and the Economic Censuses, which are fielded in years ending in 2 and 7. This information is used to maintain up-to-date company affiliation, location, and operating information for establishments that are part of multi-establishment companies in the Business Register (BR). The COS surveys all multiple-unit firms with more than 250 employees every year and smaller multiple-unit firms on a rotating basis. It maintains information on the organizational design and employment of multi-unit firms. In particular, companies identify changes in their establishments due to sale or closure, or the start-up or acquisition of new establishments. The companies are asked to indicate controlling interests held by other domestic or foreign-owned organizations. Law mandates completion of both the Economic Census and the Company Organization Survey.

The Economic Census covers all establishments of multi-unit companies, all single-unit employers larger than each industry size cutoff (for most industries about 3 employees), and a sample of small employers with fewer employees than the industry employment size cutoff in most industries except agriculture and government.²³ Its purpose is to provide comprehensive statistics about establishments and their activities. It covers all domestic non-farm business establishments, other than those operated by the government.

Main Variables

The Economic Census collects data on establishment employment in the pay period that includes March 12, total revenue, annual and first quarter payroll, full address, organizational form, and type of ownership. Additional information is collected for some sectors and industries. Those include Census of Finance, Insurance and Real Estate (CFI), Census of Manufactures (CMF), Census of Retail Trade (CRT), Census of Transportation, Communications, and Utilities (CUT), Census of Wholesale Trade (CWH), and Census of Services (CSR).

²³ See <http://www.census.gov/econ/overview/mu0000.html> (accessed June 21, 2005).

Major Limitations

A major limitation of the EC and COS is that these datasets are not publicly available. Researchers can apply to use the data at Census Research Data Centers (RDCs).²⁴ The Census Bureau releases aggregated tables from the Economic Census for industry data by firm employment size. The Census Bureau releases most tables from an Economic Census three to four years after it was conducted.

What Can be Done with the Data

As we have stated, the EC and the COS serve as two of the main surveys that add data to the BR. The type of ownership and organizational form data could be compared across firms of different size using employment-, payroll-, and revenue-based size definitions. In addition, researchers can make use of the detailed industry information collected in the economic censuses. For example, Garicano and Hubbard (2005a, 2005b) used data from the 1992 Census of Services to study specialization within and between Law firms.

SURVEY OF WOMEN/MINORITY OWNED BUSINESS ENTERPRISES (SWOBE/SMOBE), AND SURVEY OF BUSINESS OWNERS (SBO)

Data Collection Method and Coverage

Three related surveys, the samples of which are drawn from the BR data frame, are the Surveys of Minority and Women Owned Business Enterprises (SMOBE/SWOBE) and the Survey of Business Owners. These surveys are conducted once every 5 years in conjunction with the economic census. The former were carried out in 1992 and 1997 and the latter in both 1992 and 2002. The United States Code, Titles 13 and 26, authorizes these data collections and provides for mandatory responses.²⁵

These surveys supplement the BR and all related products with more detailed information on business owners. The 2002 SBO samples from all businesses that file tax forms as individual proprietorships, partnerships, or corporations with receipts of \$1,000 or more. Major industries that are excluded from the survey include agricultural production, domestically scheduled airlines, railroads, U.S. Postal Service, private households and some non-profit organizations. The sampling methodology includes an interesting method of sorting business into gender/racial/ethnic categories based on a series of probabilities obtained from additional sources.

Main Variables

These surveys collect detailed information on business owners, including age, education level, sources of financing, gender, race, and ethnic background. The 2002 SBO also collected information on veteran status, service disabled, types of customers (federal government, local government, consumers)

²⁴ For more information see <http://148.129.75.149/ces.php/guidelines>

²⁵ See <http://www.census.gov/econ/overview/mu0200.html> (accessed July 5, 2005).

and workers, home-based firms, family-owned firms, and sources of financing for capital improvements or start-up.²⁶ These data can be linked to the firm size information from the Business Register or Economic Census.

Major Limitations

Between 1992 and 1997, there were some changes in the survey methodology that might limit direct comparisons between years. In particular, changes were made to the target population (the universe was extended to include different types of corporations) and to the definition of business (in 1997 all operations under the same ownership were defined as one company even though they might have had different Employer Identification Numbers).²⁷ The Census Bureau releases detailed tables from the Survey of Business Owners approximately four years after the data is collected.

What Can be Done with the Data

These data sources provide a rich and unique set of information on the characteristics of small and large business owners and their sources of financing. One of the benefits of this survey is that it can be connected to the other census products that have longitudinal information on employment and revenues. Using the 2002 SBO one can also examine characteristics of firms that have federal and local public agencies as their main consumers.

INTEGRATED LONGITUDINAL BUSINESS DATABASE (ILBD) AND LONGITUDINAL EMPLOYER- HOUSEHOLD DYNAMIC PROGRAM (LEHD)

Data Collection Method and Coverage

Current efforts in the Bureau of Census are directed at developing integrated databases that include employer and employee characteristics, as reported by John Haltiwanger²⁸ at the Kauffman Symposium on Entrepreneurship Data held November 10-11, 2004. The focus of these efforts is on extending the LBD along two dimensions. The first extension integrates non-employer data making it possible to track transitions to and from employer to non-employer status.²⁹ The second extension is to include information from the Longitudinal Employer Household Dynamics (LEHD) files, which provide person and business identifiers for all workers and businesses covered by unemployment insurance in 30 states. The person identifiers can then be used to match workers to information in other person-level census products, such as SIPP or the Census Long Form.

²⁶ See <http://www.census.gov/econ/census02/sbo/intro.htm> (accessed June 21, 2005)

²⁷ See <http://www.census.gov/csd/mwb/comp.htm> (accessed August 25, 2005).

²⁸ Dr. John Haltiwanger is a professor of Economics at the University of Maryland and research associate at the Center for Economic Studies at the Census Bureau. He also was Chief Economist at the Census Bureau in 1997-1999.

²⁹ In 2002 there were approximately 16 million non-employer businesses. Approximately 14 million do not have an EIN but are uniquely identified by the owner social security number and the other 2 million have an EIN. (Davis et al, 2005)

The dataset including both employee and non-employee businesses is called the Integrated Longitudinal Business Database (ILBD). The ILBD combines administrative records and survey-based data for virtually all employer and non-employer business units in the U.S. The data integrate the LBD, discussed above, with the non-employer data (Davis et al., 2005). In 2000 there were roughly 14 million non-employer businesses. The census defines a non-employer business as one that has no paid employees, has annual business receipts of \$1,000 or more (\$1 or more in the construction industries), and is subject to federal income taxes.³⁰ These are usually self-employed individuals operating a very small unincorporated business. For small non-employer businesses, the Census Bureau obtains administrative records from individual income tax returns. These small non-employer businesses likely do not have an EIN, and are tracked using the self-employed individuals' Social Security Number (SSN). This new integrated data permits analysis of the movements of people between employer and non-employer universes.

The LEHD connects establishment data from LBD to household data. The employer portion of this integrated database includes records for about 4 million establishments from more than 20 states. In his presentation, Dr. Haltiwanger encouraged researchers to persevere in making use of these rich data sources. For more information see Davis et al. (2005).

Main Variables

The combined dataset includes establishment-level information on monthly employment, quarterly wages, detailed location, industry, workforce composition and worker turnover for 1990-2004 (years covered vary by state). These records are extracted from the Unemployment Insurance data and can be connected to other Census Bureau databases. In addition, many employer-level measures are created from longitudinally integrated person- and establishment-level data. The data also include (i) worker and job flows, including new hires, separations, job creation, job elimination by age and gender of workforce; (ii) worker composition by gender and age, (iii) worker compensation for stocks and flows by gender and age; and (iv) dynamic worker compensation summary statistics for stocks and flows by gender and age. This file also contains identifiers used by state ES-202 and Unemployment Insurance systems as well as federal employer IDs (EINs). A description of the data is presented in LEHD (2002).

Major Limitations

A major limitation of this dataset is that it is not publicly available. Researchers can apply for access to the microdata at the Census Research Data Centers.³¹

³⁰ See definition at <http://www.census.gov/epcd/nonemployer/view/define.html> (Accessed June 21, 2005).

³¹ For detailed description see <http://www.ces.census.gov/ces.php/rdc> (accessed June 20, 2005).

What Can be Done with the Data

This dataset has great potential for studying dynamic changes in establishments and firms and connecting these to owner and worker characteristics. The dataset includes more data elements than found in business owner surveys and covers an important part of the small business universe - non-employer businesses. Current research is using the LEHD to study the impacts of new technologies on firms and workers (Abowd, et al, 2001); measure the relationships between human capital and a firm's technology (Abowd, et al, 2004); and examine the relationship between employer provided health insurance, worker mobility and wages (Stinson, 2003). For other papers that used LEHD see U.S. Census Bureau (2004).

OTHER GOVERNMENT SOURCES OF DATA

SURVEY OF SMALL BUSINESS FINANCES

Data Collection Method and Coverage

In addition to these Census and BLS sources of data, the Federal Reserve Board sponsors a survey of small firms. The survey of Small Business Finances (SSBF) was conducted in 1987, 1993, 1998, and 2003. The SSBF contains information from more than 3,500 firms with fewer than 500 employees. It oversampled African-American, Asian-American, and Hispanic-American-owned firms. The sampling frame for the SSBF is the Dun's Market Identifier (DMI) file, which is described below. Unlike most of the government data sources described here, full public datasets for each of the first three years of SSBF are available from the Federal Reserve. This survey appears to have been well designed and the sampling plan and implementation are well documented.

Main Variables

This source includes information on firm employment, owner characteristics, use of financial services, and the income and balance sheets of the firm. For example, the survey has characteristics of the profits, sales, labor productivity, or other measures of firm success. It also collects information on the experience of small businesses with credit applications and credit access.

Major Limitations

Users of the survey should keep in mind that the survey was voluntary, and the response rate was approximately thirty-three percent. The survey also provides no information at the establishment level.

What Can be Done with the Data

Researchers have used this survey to examine financial constraints that firm face (Lel and Udell, 2002; Robb, 2002); adoption of computers (Bitler, 2001); use of financial services (Bitler, Robb, and Wolken, 2001); borrowing experience by gender, race and ethnicity of firm owners (Coleman, 2002a; Coleman, 2002b; Coleman, 2003); and the decision to become a public firm (Helwege and Packer, 2003). A description of the studies that used SSBF is provided at the Federal Reserve website.³²

NATIONAL EMPLOYER HEALTH INSURANCE SURVEY

Data Collection Method and Coverage

A number of federal agencies collect data related to the provision of health insurance. For instance, in 1994 the Center for Disease Control and Prevention conducted the National Employer Health Insurance

³² See <http://www.federalreserve.gov/pubs/oss/oss3/abstract.html> (accessed June 21, 2005).

Survey (NEHIS). The survey was designed to produce estimates on employer-sponsored health insurance data in the United States for establishments of different sizes.³³

The NEHIS is a 1994 national probability sample survey of establishments, public sector entities, and self-employed individuals.³⁴ The sample was drawn from three sample frames – The Dun’s Market Identifiers file for the sample of private establishments, the Census of Governments file for public sector entities, and the National Health Interview Survey to select self-employed individuals with no employees (NCHS, 1997). Among private sector establishments, 34,604 interviews were completed for a response rate of 71 percent.

Main Variables

The NEHIS collected detailed data on various aspects of employer-sponsored health insurance, types of plans offered to employees, and detailed characteristics of the offered plans (monthly premiums, employer and employee contributions to premiums, deductibles, coinsurance, and covered services). The survey also collected information on establishment’s employment, industry, and location.

Major Limitations

Due to confidentiality restrictions, this survey is available to researchers only through the National Center for Health Statistics Research Data Center (RDC).³⁵ Prospective researchers must submit research proposals to RDC. In addition, there are costs associated with working in the center - \$500 per month for remote access and \$1000 per week of on site access.

MEDICAL EXPENDITURES PANEL SURVEY

Data Collection Method and Coverage

The Medical Expenditures Panel Survey (MEPS), conducted by the Agency for Healthcare Research and Quality, also contains information on the health insurance offerings of businesses of different employment sizes. The MEPS includes four components that provide an important overview of access to health insurance and care. It includes a Household Component (HC), Nursing Home Component (NHC), Medical Provider Component (MPC), and Insurance Component (IC). The Household and Insurance Components provide some information valuable to small businesses researchers. The MEPS collects data on health services used by Americans, frequency of use, cost of services, and method of payment. In addition, data is collected on the cost, scope, and breadth of private

³³ See description on <http://www.cdc.gov/nchs/about/major/nehis/nehis.htm> (accessed June 21, 2005)

³⁴ See description of the methodology on <http://www.cdc.gov/nchs/about/major/nehis/method.htm> (accessed June 21, 2005).

³⁵ See <http://www.cdc.gov/nchs/about/major/nehis/pudf.htm> (accessed June 21, 2005).

health insurance held by and available to the U.S. population.³⁶ Yearly surveys have been conducted since 1996.

The HC collects data from a sample of families and individuals around the nation. The sample is drawn from the nationally representative sub-sample of households that participated in the prior year's National Health Interview Survey. The households are interviewed for five rounds that cover two full calendar years.

The IC derives information from two subcomponents: the household sample and the list sample. The "household sample" includes employers and other insurance providers (unions and insurance companies) of respondents to the previous year's MEPS HC. The number of respondents in this sample varies from year to year. These data, when linked back to the original household respondent, allow for the analysis of individual behavior and choices made with respect to health care use and spending.³⁷ The list sample includes business establishments and governments. The sample frame is derived from the BR. The target size of the list sample is approximately 30,000 per year.

Main Variables

The Household Component survey collects data on the size of employer, type of industry, wage level, weekly hours of work, and current employment status (working, unemployed, or retired). In addition, it tracks changes in respondents' health status, income, employment, eligibility for public and private insurance coverage, use of services, and payment for care.

The insurance component collects information on firm total employment, employment at chosen establishment, number of employees eligible for coverage, employment characteristics, and full address. It also collects information on the number of health insurance plans that were offered by the establishment, as well as other establishment characteristics.

Major Limitations

There are several limitations to using this dataset. The questionnaire and coverage have changed somewhat over time. For example, information from the sample of self-employed workers was collected only in 1996 and was discontinued after that. New questionnaires were added in 1997 for cases in which the collection took place at company headquarters for ten or more establishments. In 1998, additional questionnaires were added to deal with follow-up interviews for multi-unit establishments at the company level.

In addition, due to confidentiality constraints, MEPS microdata can be accessed only through the Data Centers at the Agency for Healthcare Research and Quality.

³⁶ See description of the program at <http://www.meps.ahrq.gov/> (accessed June 21, 2005).

³⁷ See description on http://www.meps.ahrq.gov/Data_Pub/Questionnaires/questic/readmeic.htm (accessed June 21, 2005).

What Can be Done with the Data

Several studies have used MEPS. For example, Gresenz, Rogowski, and Escarce (2004) used the data to study access to care among the uninsured. This dataset can also potentially be used to study characteristics of firms that provide health insurance coverage for workers, as well as responses of workers to different types of insurance provided by the firm (e.g. the selection of workers into different firms based on the health insurance coverage, or compensation for workers who secure health insurance from other sources). Some of the component datasets allow for analysis of the self-employed, costs of providing insurance for small firms, and the impact of changes in federal and state health care policies.

PRIVATE AND COMMERCIALY AVAILABLE DATA SOURCES

Although government administrative datasets and surveys provide a wealth of information on small businesses, there are two important reasons for also considering private data sources. The first is that few of the government data sources, and neither the longitudinal data sources nor the potential data frames of all businesses are publicly available. The second reason is that, as with all data sources, the information collected by government sources typically seeks to answer a particular question or mandate and may not include the information necessary to address other timely policy questions on particular populations of interest.

The private data sources described here are publicly available but may carry considerable cost. In addition, the needs of researchers may not have been the primary concern for those collecting the data. Thus, these sources can raise a variety of other data issues such as coverage, representativeness, and for survey sample sources, response rates.

DUNS MARKET IDENTIFIER (DMI)

Data Collection Method and Coverage

The private data source most widely used by both governmental and private organizations for research on small businesses is Dunn and Bradstreet's list of U.S. businesses, the Duns Market Identifier (DMI) file. While the Census Bureau and the Bureau of Labor Statistics use separate master files of U.S. businesses as sampling frames for their own surveys, they do not make these lists available to other governmental agencies or private organizations due to confidentiality concerns. For these agencies and organizations, the DMI is currently the most complete listing of businesses from which to draw a potentially nationally representative probability sample.

The DMI contains information on more than 14 million U.S. businesses. DMI data are continually updated based on information from a variety of sources—primarily D&B's credit rating service and business directories—but also from direct investigation and interviews, payment and banking data from company suppliers, suits, liens, judgments, business registrations, bankruptcy filings, corporate financial reports, government contracts, grants, loans and debarments, data mining of more than 27 million internet domains, news and media sources, and yellow page and print directories. The frequency of information updates for a business is a function of its employment, industry, and activity level.

Main Variables

The DMI includes measures of firm and establishment employment (employment includes owners or unpaid family members who are workers, which differs from the definition used by government data sources), and annual sales. In addition, the listing provides detailed location data: telephone number,

location, and owner name. Data are also collected for owner minority status, industrial classification, firm's start year, and legal status.

Major Limitations

The DMI is neither longitudinal nor does it provide a snapshot of information from a particular time point since data are updated at different intervals for each business. It is difficult to gauge the range of DMI coverage, especially of new and small businesses. Over time D&B has expanded its coverage of small businesses and small establishments within larger firms. Several recent surveys that used the DMI as a sampling frame found it useful to first screen selected businesses to make sure that they were currently active and were private firms with employees. The National Employer Health Insurance Survey, for instance, found that approximately 18 percent of the screened businesses did not meet these criteria. The Kauffman Firm Survey (described below) will soon be able to provide more up-to-date information on the possible over- or under-coverage of the DMI.

What Can be Done with the Data

The data is most commonly used as a sampling frame for new surveys of firms. DMI data are available for purchase from Dunn and Bradstreet, which makes the listing more accessible than other sources.

KAUFFMAN FIRM SURVEY (KFS)

Data Collection Method and Coverage

The Kauffman Firm Survey (KFS) is a new survey commissioned by the Kauffman Foundation to provide publicly available longitudinal data on new firms. Researchers currently have completed two rounds of pilot data collection, and the first full panel is scheduled to go into the field the summer of 2005. The goal of the KFS is to longitudinally track new firms over time with an emphasis on financial development, high technology and women-owned firms. Two cohorts of new firms will each be followed for multiple years. The first cohort of approximately 5,000 firms, new in 2004, will be followed for three additional years and the second panel of 5,000 firms, new in 2006, for one additional year. The KFS is using listings from the DMI with a 2004 start year as the sampling frame for the first cohort and oversampling high-technology and women-owned firms. Although details are not yet available, the expectation based on the pilot data is that only 40 percent of the firms selected will be eligible for the study due to errors in DMI data or over-coverage of the frame and that of these, approximately 40 percent will complete the survey.

The first stage of sampling includes a 10-item screener for "new firms". This screener includes questions about the timing of first paying unemployment insurance taxes, payment of social security taxes, submission of a Schedule C for business income or losses, and application for an EIN. This

information could be used to determine when new firms are recorded in government data sources (BITS and BED) that use one or more of these indicators to mark firm births. In addition, the screener collects information on possible forms of a business' legal status (there are seven options including sole proprietorship, limited partnership, and limited liability company).

Main Variables

The full survey collects information about firm employment, including number of full and part-time workers. In addition, the survey collects information about the owner's work behavior and demographics, characteristics of the firm, business strategy and use of innovation, business organization, human resources benefits, and detailed information on finances.

Major Limitations

This survey is not yet fielded and predicts a low response rate.

RESEARCH DATASET DERIVED FROM THE MARTINDALE-HUBBELL LAW DIRECTORY

Data Collection Method and Coverage

In recent years, researchers have synthesized databases from various listings and directories. One example is a research database of law firms extracted from the Martindale-Hubbell Law Directory. Martindale Hubbell (MH) is the leading reference on the American legal-services industry. The MH publishes listings for almost the entire universe of lawyers and law firms, but the underlying data for each firm are not available. However, researchers can use directory listings to synthesize the database.

Romley and Talley (2004) extracted the listings of all lawyers and law firms from the HM law directory for 1993 and 1999. A matching mechanism was used to create a firm-level database with information on each establishment within a firm. There are approximately 65,000 law firms in each year and firms are connected between years.

Main Variables

The database includes information on the number of offices within a firm, office- and firm-level employment, distribution of employment by type of position, quality rankings, types of law fields in which the firm operates, and organizational form.

Major Limitations

There are some limitations to this constructed database. First, this research database is not publicly available, although researchers can design automated algorithms that would re-synthesize the database. Second, the current matching algorithm is limited. The algorithm matches law offices to a law firm based on the name only. This can lead to false matching of unrelated law firms that have the same name. This

problem is especially pronounced in the case of small firms that have the lawyer's name as firm name. The algorithm can be improved by allowing matching also by year of the firm's inception. In addition, so far this research database covers only two years. Additional years of the data can be extracted.

What Can be Done with the Data

Romley and Talley (2004) used this longitudinal dataset to examine effects of the availability of new organizational forms for law firms of different employment sizes. More generally, researchers can use this dataset to examine the effects of different policies on access to and quality of lawyers as well as their specialization within the legal profession.

THE KAISER FAMILY FOUNDATION/HEALTH RESEARCH AND EDUCATIONAL TRUST EMPLOYER HEALTH BENEFITS SURVEYS

Data Collection Method and Coverage

There are also private databases with information about small firms and health insurance offerings. One such data source is the Kaiser Family Foundation/Health Research (Kaiser/HRET) Employer Health Benefits Surveys, which have been fielded annually since 1999. Previous versions of the survey were sponsored by the Health Insurance Association of America from 1987–1990 and KPMG from 1991–1998. The nationally representative sample was drawn from the Dun & Bradstreet list of the nation's employers with three or more workers and stratified by firm employment. Approximately 50 percent of the sample members in 2004 were participants in either the 2002 or 2003 survey, so some panel data are available. The survey covers about 3,000 employers each year.

Main Variables

The survey collects information on the number of workers in a firm, health insurance offerings, the cost of coverage, and health insurance attitudes and opinions.

Major Limitations

Response rates for this survey are not made publicly available and requests for data are handled on an individual basis.

DISCUSSION

Great strides have been made in recent years to create data sources useful for conducting research on policy and small business. Of particular importance are new longitudinal datasets created by both the Census Bureau and the Bureau of Labor Statistics, which allow changes within establishments and firms to be studied over time.

In creating administrative longitudinal databases, progress was made on three problematic issues: connecting establishments to parent firms, matching establishments and firms over time, and identifying

firm start dates and closures. This substantial work made it possible to create the datasets described above; however, challenges in using these data remain. Researchers need to carefully consider how well the three issues were addressed in each data source and whether the resulting quality of the data might impact their research.

The most notable gap in current small business data sources is the lack of a *publicly available* source of longitudinal data. In the next five years, this gap will be at least partially addressed by the Kauffman Firm Survey of new businesses. Moreover, while the BITS and BED databases are not publicly available, individuals at the Census Bureau have voiced interest in more researchers making use of them. In his presentation at the Kauffman Symposium on Entrepreneurship Data (November, 2004), Dr. Haltiwanger strongly encouraged researchers both to submit proposals for using the data at Census Research Data Centers and to continue to support the Census Bureau's use of the BED as a frame for survey samples through which more detailed and specific information can be collected and made publicly available.

KEY INFORMATION SOURCES FOR SMALL BUSINESS DATA

GENERAL RESOURCES ON DATA FOR SMALL BUSINESSES

History of Government Small Business Data Collection Efforts:

Armington, C. Development of Business Data: Tracking Firm Counts, Growth, and Turnover by Size of Firms, Washington, D.C.: SBA Office of Advocacy, December 2004. Online at <http://www.sba.gov/advo/research/rs245tot.pdf>

Current Research and Resources from the Small Business Administration, Office of Advocacy:

SBA Office of Advocacy, Research Publications 2004. Online at http://www.sba.gov/advo/research/res_pub04.pdf, (accessed July 5, 2005).

The Small Business Economy: A Report to the President, 2004 Edition, Editor Kathryn Tobias, SBA Office of Advocacy. Online at http://www.sba.gov/advo/stats/sb_econ2004.pdf

SBA list of links to resources on small business data

This is a good fairly complete list of government data sources on small business. Online at <http://app1.sba.gov/faqs/faqindex.cfm?areaID=2>

CENSUS BUREAU DATA SOURCES

Research using BITS:

Acs, Z. and Armington, C. Using Census BITS To Explore Entrepreneurship, Geography, and Economic Growth, Ruxton, MD: SBA Office of Advocacy, February 2005. Online at <http://www.sba.gov/advo/research/rs248tot.pdf> (accessed July 5, 2005).

National Women's Business Council. "Firms Owned By Women of Color Show Staying Power: African American Firms Lag Somewhat Behind," Washington, D.C.: August 24, 2004. Online at http://www.nwbc.gov/documents/FINAL-Census-Bureau_Trends08.24.04.pdf

National Women's Business Council. "Women Employer Firms Continue to Show Strength: Similar Survival Rates, Fewer Job Losses From 1997–2001," Washington, D.C.: February 14, 2005. Online at http://www.nwbc.gov/NewsCenter/documents/final_census_bureau_trends_news_release_02-09-05.pdf

Robb, A. New Data for Dynamic Analysis: The Business Information Tracking Series (BITS), SBA Office of Advocacy. Online at http://permanent.access.gpo.gov/lps4810/bits_doc.pdf

Venegas, E. Issue Brief: Start-ups and Expansions Fuel Minnesota's Economic Engine, St. Paul, MN: Minnesota Department of Trade and Economic Development, December 2000. Online at <http://www.deed.state.mn.us/bizdev/PDFs/startups&expan.pdf>

Villalonga, B. Diversification discount or premium? New evidence from BITS establishment-level data, Los Angeles: The Anderson School at UCLA, November 2000. Online at <http://148.129.75.160/paper.php?paper=101628>

Information on the 2002 Economic Census

U.S. Census Bureau website, 2002 Economic Census, reports, data tables, links to other economic census years. Online at <http://www.census.gov/econ/census02/>

Information on the SWOBE/SMOBE/SBO (Surveys of Women and Minority Owned Business Enterprises/ Survey of Business Owners)

U.S. Census Bureau website, 2002 Economic Census, report on the Survey of Business Owners: Advance Report on Characteristics of Employer Business Owners. Online at <http://www.census.gov/econ/census02/sbo/intro.htm> (accessed July 5, 2005)

U.S. Census Bureau website, 1997 Economic Census Surveys of Minority- and Women- Owned Business Enterprises, links to surveys, reports, data. Online at <http://www.census.gov/csd/mwb/>

Information on the Company Organization Survey

U.S. Census Bureau website, Company Organization Survey, overview and brief methodology description, links to County Business Patterns. Online at <http://www.census.gov/econ/overview/mu0700.html>

BUREAU OF LABOR STATISTICS DATA SOURCES

Research using BED:

Spletzer, J., Faberman, R., Sadeghi, A., Talan, D., and Clayton, R. "Business employment dynamics: new data on gross job gains and losses," Monthly Labor Review, April 2004. Online at <http://www.bls.gov/opub/mlr/2004/04/art3full.pdf>

U.S. Department of Labor, Bureau of Labor Statistics website, links to statistical resources and publications on business employment dynamics. Online at <http://www.bls.gov/bdm/home.htm#publications>

U.S. Department of Labor, Bureau of Labor Statistics website, links to state websites containing geocoded QCEW research. Online at <http://www.bls.gov/bdm/geocode.htm>

Research using the CES (Current Employment Statistics)

U.S. Department of Labor, Bureau of Labor Statistics website. Employment, Hours, and Earnings from the Current Employment Statistics survey. Online at <http://www.bls.gov/ces/home.htm#publications>

Research on Small Businesses using the CPS (Current Population Survey):

SBA Office of Advocacy website, Characteristics of Small Business Employees and Owners, 1997. A reference guide on the workers and owners of small businesses prepared by the Office of Economic Research of the U.S. Small Business Administration's Office of Advocacy. Online at http://www.sba.gov/ADVO/stats/ch_emp_o.html#2

Data on Non-employer as well as Employer Businesses – the Integrated Longitudinal Business Database (ILBD)

"Measuring the Dynamics of Young and Small Businesses: Integrating the Employer and Non-employer Universes" by Steven J. Davis, John Haltiwanger, Ron Jarmin, C.J. Krizan, Javier Miranda and Alfred Nucci and Kristin Sandusky. 2005. Online at http://www.aeaweb.org/annual_mtg_papers/2005/0107_0800_0403.pdf

OTHER GOVERNMENT DATA SOURCES

Information on the SSBF including methodology reports, papers, codebooks, survey instruments and data

Federal Reserve Board website, Survey of Small Business Finances. Links to working papers and methodology reports, codebooks and other related documentation, and the full public datasets are

available for the 1998, 1993, and 1987 SSBF. A list of answers to frequently asked questions (FAQ) is also available. Online at <http://www.federalreserve.gov/pubs/oss/oss3/nssbftoc.htm>

Information on the IRS Survey of Income -related research projects

Internal Revenue Service website, Survey of Income, Projects and Contacts. Online at <http://www.irs.gov/pub/irs-soi/projcont.pdf>

PRIVATE SOURCES OF DATA

Dunn & Bradstreet

Description of the system used to insure data quality online at http://www.dnb.com/us/about/db_database/dnbinfoquality.html

Standard and Poor's Compustat data resources for research

Description of software and ordering information online at http://www.compustat.com/www/ug/univ_mkt.html

Information on the 2004 Kaiser-HRET Employer Health Benefits Survey

This annual survey of employers provides detailed insights into trends in employer-based health coverage, including changes in premiums, employee contributions, cost-sharing policies and other relevant information. Online at <http://www.kff.org/insurance/7148/index.cfm>

CONFERENCE RESOURCES:

Entrepreneurship in the 21st Century (March 26, 2004), Conference Proceedings

http://www.sba.gov/advo/stats/proceedings_a.pdf

http://www.sba.gov/advo/stats/proceedings_b.pdf

http://www.sba.gov/advo/stats/proceedings_c.pdf

Kauffman Symposium on Entrepreneurship Data (Nov 10-11, 2004)

Conference Summary Document available by request from A. Haviland

NAS PANELS:

Measuring Business Formation, Dynamics and Performance. Links to Project Scope and panel membership available at http://www4.nas.edu/cp.nsf/Projects+_by+_PIN/CNST-I-04-01-A?OpenDocument

Review of Federal Business Statistics, National Academies, Division of Behavioral and Social Sciences Education, Committee on National Statistics. Link to latest meeting agenda <http://www4.nas.edu/webcr.nsf/MeetingDisplay1/CNST-I-04-01-A?OpenDocument>. Panel Members: John Haltiwanger (co-chair), Lisa Lynch (co-chair), John Abowd, Patricia Anderson, Matthew Barnes, Steven Davis, Timothy Dunne, Robert Groves, Susan Hanson, Robert McGuckin, Paul Reynolds, Mark Roberts, Niels Westergaard-Nielsen, Kirk Wolter. Final Report will be issued January 2006.

REFERENCES

- Abowd, John, John Haltiwanger, Julia Lane, and Kristin Sandusky, *Within and Between Firm Changes in Human Capital, Technology, and Productivity*, 2001.
- Abowd, John, John Haltiwanger, Ron Jarmin, Julia Lane, Paul Lengermann, Kristin McCue, Kevin McKinney, and Kristin Sandusky, *The Relation among Human Capital, Productivity and Market Value: Building Up from Micro Evidence*, August 1, 2004.
- Acs, Zoltan J., and Catherine Armington, *Longitudinal Establishment And Enterprise Microdata (LEEM) Documentation*, CES-WP-98-9 May 1998.
- Armington, Catherine, Alicia Robb, and Zoltan J. Acs, *MEASURES OF JOB FLOW DYNAMICS IN THE U.S.*, CES-WP-99-1 January 1999.
- Armington, Catherine, and Alicia Robb “Mergers and Acquisitions in the United States: 1990-1994,” CES-WP-98-15, September 1998.
- Barsky, Carl B., “Incidence Benefits Measures in the National Compensation Survey,” *Monthly Labor Review*, Vol. 127, No. 8, August 2004, pp. 21-28.
- Berger, Mark C., Dan A. Black, Frank A. Scott, and Steven N. Allen, *Distribution of Low-Wage Workers by Firm Size in the United States*, SBA publication # 196, March 2000.
- Bernstein, Jared, and Maury Gittleman, “Exploring Low-Wage Labor with the National Compensation Survey,” *Monthly Labor Review*, Vol. 126, No. 11-12, 2003, pp. 3-12.
- Bitler, Marianne, “Small Businesses and Computers: Adoption and Performance,” Federal Reserve Bank of San Francisco, Working Paper: 2001-15.
- Bitler, Marianne, Alicia M. Robb, and John D. Wolken, “Financial Services Used by Small Businesses: Evidence from the 1998 Survey of Small Business Finances,” *Federal Reserve Bulletin*, Vol. 87, No. 4, 2001, pp. 183-205.
- Blostin, Allan P., “The National Compensation Survey: A wealth of benefits data,” *Monthly Labor Review*, 2004, online at http://www.findarticles.com/p/articles/mi_m1153/is_8_127/ai_n6358418/print (accessed June 9, 2005)
- Carrington, William J., and Kenneth R. Troske, “Gender Segregation in Small Firms,” *The Journal of Human Resources*, Vol. 30, No. 3, 1995, pp. 503-533.
- Coleman, Susan, “Borrowing Patterns for Small Firms: A Comparison by Race and Ethnicity,” *The Journal of Entrepreneurial Finance & Business Ventures*, Vol. 7, No. 3, 2003, pp. 87-108.
- Coleman, Susan, “Characteristics and Borrowing Behavior of Small, Women-Owned Firms: Evidence from the 1998 Survey of Small Business Finances,” *The Journal of Business and Entrepreneurship*, Vol. 14, No. 2, 2002b, pp. 151-166.
- Coleman, Susan, “The Borrowing Experience of Black and Hispanic-Owned Small Firms: Evidence from the 1998 Survey of Small Business Finances,” *The Academy of Entrepreneurship Journal*, Vol. 8, No. 1, 2002a, pp. 1-20.

- Copeland, Kennon, Nonresponse Adjustment in the Current Employment Statistics Survey, Federal Committee on Statistical Methodology, 2003, online at www.fcsm.gov/03papers/Copeland.pdf (accessed July 5, 2005).
- Diamond, Charles A., and Curtis J. Simon, "Industrial Specialization and the Returns to Labor," *Journal of Labor Economics*, Vol. 8, No. 2, 1990, pp. 175-201.
- Dietz, Elizabeth, "Trends in Employer-Provided Prescription-Drug Coverage," *Monthly Labor Review*, Vol. 127, No. 8, August 2004, pp. 37-45.
- Evans, David S., and Linda S. Leighton, "Empirical Aspects of Entrepreneurship," *The American Economic Review*, Vol. 79, No.3, 1989, pp. 519-535.
- Fairlie, Robert, Technology and Entrepreneurship: A Cross-Industry Analysis of Access to Computers and Self-Employment, SBA Report # 259, 2005.
- Foster, Lucia, Establishment and Employment Dynamics in Appalachia: Evidence from the Longitudinal Business Database, CES-WP-03-19 December 2003, online at <http://www.ces.census.gov/ces.php/abstract?paper=101690> (accessed August 25, 2005).
- Garicano, Luis, and Thomas N. Hubbard, Learning About the Nature of Production from Equilibrium Assignment Patterns, April 25, 2005, online at <http://gsbwww.uchicago.edu/fac/thomas.hubbard/research/> (accessed July 5, 2005).
- Garicano, Luis, and Thomas N. Hubbard, Managerial Leverage Is Limited By the Extent of the Market: Hierarchies, Specialization, and the Utilization of Lawyers' Human Capital, January 2005, online at <http://gsbwww.uchicago.edu/fac/thomas.hubbard/research/> (accessed July 5, 2005).
- Garicano, Luis, and Thomas N. Hubbard, Specialization, Firms, and Markets: The Division of Labor Within and Between Law Firms, May 2005, online at <http://gsbwww.uchicago.edu/fac/thomas.hubbard/research/> (accessed July 5, 2005).
- Gresenz, Carole Roan, Jeannette A. Rogowski, and José J. Escarce, Health Care, Markets, the Safety Net and Access to Care Among the Uninsured, NBER Working Paper 10799, 2004, online at <http://www.nber.org/papers/w10799> (accessed June 22, 2005).
- Headd, Brian, "Business Success: Factors Leading to Surviving and Closing Successfully," CES-WP-01-01 January 2001.
- Helwege, Jean and Frank Packer, "The Decision to Go Public: Evidence from Mandatory SEC Filings of Private Firms," Fisher College of Business Working Paper. Ohio State University April 2003.
- Idson, Todd L., and Daniel J. Feaster, "A Selectivity Model of Employer-Size Wage Differential," *Journal of Labor Economics*, Vol. 8, No. 1, 1990, pp. 99-122.
- Jarmin, Ron S., Shawn D. Klimek, Javier Miranda, Firm Entry and Exit in the U.S. Retail Sector, 1977-1997, CES-WP-04-17 October 2004, online at <http://www.ces.census.gov/ces.php/abstract?paper=101704> (last accessed August 25, 2005).
- Jarmin, Ron S., and Javier Miranda, The Longitudinal Business Database, 2002, online at <http://www.ces.census.gov/paper.php?paper=101647&PHPSESSID=838fec06e9652892337cde662d45ed5a> (last accessed June 20, 2005)

Karoly, Lynn A., and Julie Zissimopoulos, *Self-Employment Trends and Patterns Among Older U.S. Workers*, RAND Working Paper #136, 2003.

LEHD program, The Longitudinal Employer-Household Dynamics Program: Employment Dynamics Estimates Project Versions 2.2 and 2.3, LEHD Technical Paper 2002-05, 2002, online at <http://lehd.dsd.census.gov/led/library/techpapers/tp-2002-05-rev1.pdf> (accessed June 21, 2005).

Lel, Ugur and Gregory F. Udell, "Financial Constraints, Start-up Firms and Personal Commitments." Kelley School of Business Working Paper. Indiana University. October 2002.

Madrian, Brigitte C., and Lars John Lefgren, A Note on Longitudinally Matching Current Population Survey (CPS) Respondents, NBER Technical Working Paper 247, Cambridge, MA: National Bureau of Economic Research, November 1999, online at <http://www.nber.org/papers/T0247>, (accessed June 9, 2005).

National Center for Health Statistics, Employer-sponsored health insurance: State and national estimates, Hyattsville, Maryland, 1997, online at <http://www.cdc.gov/nchs/products/pubs/pubd/other/miscpub/nehisrev.htm> (accessed June 21, 2005).

Robb, Alicia, "Small Business Financing: Differences Between Young and Old Firms," *Journal of Entrepreneurial Finance and Business Ventures*, (November 2002).

Robb, Alicia, NEW DATA FOR DYNAMIC ANALYSIS: THE LONGITUDINAL ESTABLISHMENT AND ENTERPRISE MICRODATA (LEEM) FILE, CES-WP-99-18 December 1999.

Romley, John and Talley, Eric L., "Uncorporated Professionals," USC Law and Economics Research Paper No. 04-22, and USC CLEO Research Paper No. C04-18, September 3, 2004, online at <http://ssrn.com/abstract=587982> (accessed June 22, 2005).

Stinson, Martha, Estimating the Relationship between Employer-Provided Health Insurance, Worker Mobility, and Wages, U.S. Census Bureau, LEHD Program Technical paper No. TP-2002-23

The Bureau of Labor Statistics, "The BLS Handbook of Methods", April 1997, online at <http://www.bls.gov/opub/hom/homtoc.htm> (accessed June 14, 2005).

The U.S. Department of Labor Bureau of Labor Statistics, and Census Bureau, Current Population Survey: Design and Methodology, Technical Paper 63RV, 2002, online at <http://www.census.gov/prod/2002pubs/tp63rv.pdf> (accessed June 17, 2005).

U.S. Census Bureau, LEHD Program, LEHD Bibliography, Technical paper, March 29, 2004

APPENDIX A

DESCRIPTION OF MAIN PARTS OF THE TABLE

Collection method and coverage: information about the origin and sources of the data. Note that some of the entries refer to tables aggregated from the actual dataset. The number of observations includes the most recent date for which the data are reported unless stated otherwise.

Topic, main variables: the most important variables in the dataset as well as information on the most precise geographical identifiers available (region/state/county/zip code/city/full address).

Periodicity and dates available: range of years for which data are available. This column also indicates whether the data are cross-sectional or whether the underlying units of observations are connected over time into a panel.

Unit of observation: units for which data are gathered and any linkages to higher levels of observation (e.g., establishment-level data aggregated at the level of firm).

Employment definition: typically defined as the number of people on payroll in the pay period that includes the 12th of the month.

Major limitations: –the main limitations of the dataset.

Table 1
Summary of the available datasets

Name of the dataset; Source	Collection method and coverage	Topic, Main Variables	Periodicity, Dates available, longitudinal links	Unit of observation	How employee is defined	Major limitations
Quarterly census of Employment and Wages (QCEW), also known as ES-202; BLS	Administrative records. Includes all establishments covered by UI and UCFE, about 8.4 million establishments	Employment, wages, full address (both mailing and physical location)	Quarterly, cross- section; 2001 forward (NAICS basis); 1975-2000 (SIC basis)	Establishment, can be aggregated to the firm level using EIN	Everyone on payroll	Not publicly available.* Some aggregated tables are available. ⁺ Excludes self- employed, unpaid family members, elected officials. UI coverage is different by state.
Business Employment Dynamics (BED); BLS	Connects 6.4 million establishments from QCEW over time using SESA-ID and probability matching	Monthly employment, wages, job gains and losses, full address	Quarterly, panel, 1992-forward	Establishment, can be aggregated to the firm level using EIN	Everyone on payroll	Not publicly available.* Some aggregated tables are available. ⁺ Excludes government employees, private households, and establishments with zero employment. UI coverage differs by state and may change over time.
Current Employment Statistics Survey (CES); BLS	Monthly sample survey of about 160,000 businesses and government agencies covering about 400,000 establishments	Employment, hours, and earnings, industry detail, full address	Monthly, cross- section, 1990- ongoing, some series are available since 1939	Establishment	Everyone on payroll	Not publicly available.* Some aggregated tables are available. ⁺ Establishments are not connected over time. Non-response.
National Compensation Survey (NCS); BLS	Survey , sampling frame is QCEW, three-stage design: regions, establishments, and occupations. About 19,000 establishments. Large firms are more likely to be selected.	Benefits and wages, firm employment	Yearly, cross- section	Occupations within an establishment	Everyone on payroll	Not publicly available.* Change in methodology in 1999.

*Researcher can apply for the access to the confidential micro data. For details see <http://www.bls.gov/bls/blsresda.htm> (accessed June 8, 2005).

⁺See for example, <http://www.bls.gov/cew/> (accessed June 8, 2005).

⁺See for example, <http://www.bls.gov/bdm/> (accessed June 8, 2005).

⁺See for example, <http://www.bls.gov/ces/> (accessed June 8, 2005).

⁺Researchers can apply for access to confidential data <http://www.ces.census.gov/ces.php/rdc> (accessed June 9, 2005)

Name of the dataset; Source	Collection method and coverage	Topic, Main Variables	Periodicity, Dates available, longitudinal links	Unit of observation	How employee is defined	Major limitations
Current Population Survey (CPS); BLS, Census Bureau	Monthly sample survey of approximately 60,000 households. Rotating sample design, respondents are in for 4 months, out for 8 month and in for an additional 4 months.	Firm employment, business ownership, self-employment, some characteristics of Small Business employee	Monthly, 1962-ongoing. Respondents are in for 4 months, out for 8 months and in for an additional 4 months.	Household, family, person	Not explicitly discussed	Categorical size count, matching over time is imperfect
Standard Statistical Establishment Listing (SSEL) or Business Register (BR); Census Bureau	List of all establishments and companies with paid employees; 180,000 multi-unit companies, representing 1.5 million affiliated establishments, 5 million single-establishment companies, and nearly 14 million non-employer businesses. Administrative data from IRS and SSA. Also compiles data from economic censuses and current business surveys.	Employment, revenues, business full address, organization type, industry classification, operating data, EIN.	Yearly, cross-section, 1974-2001	Establishment, and firm	NA	Not publicly available.
Longitudinal Business Database (LBD); Census Bureau	Matched records from SSEL over time using PPN, or CFN and EIN, or using name and address match. Covers all non-farm private economy and some public sector activities. 4.5-7.1 million records per year.	Establishment age and tenure, payroll, employment, firm affiliation, full address	Yearly, panel, 1974 – 1999, ongoing	Establishment, firm	Everyone on payroll	Not publicly available.▲
Business Information Tracking System (BITS), also known as Longitudinal Establishment and Enterprise Microdata (LEEM); Census Bureau	Links SSEL establishments over time using PPN, CFN, or EIN. Includes establishments with positive payroll. 13 million establishments. Same industry coverage as CBP.	Employment, firm employment, payroll, firm ownership, firm affiliation, census geography, primary industry, starting year, census file number	Yearly, panel, 1989-ongoing	Establishment, firm	Everyone on payroll	Not publicly available.▲

▲Researchers can apply for access to confidential data <http://www.ces.census.gov/ces.php/rdc> (accessed June 9, 2005)

Name of the dataset; Source	Collection method and coverage	Topic, Main Variables	Periodicity, Dates available, longitudinal links	Unit of observation	How employee is defined	Major limitations
County Business Patterns (CBP); Census Bureau	Aggregated tables derived from SSEL, excludes some agriculture, rail transportation; private households; and public administration.	Employment, payroll, total number of establishments, county	Yearly, cross-section, 1977 – forward	Establishment, firm	Everyone on payroll	Aggregated tables, some industry level data is not disclosed
Integrated Longitudinal Business Database (ILBD); Census Bureau	Connects establishment data from LBD to statistics of non-employers	See LBD	Yearly, panel, 1992-2001	Establishment	Everyone on payroll	Not publicly available.▲
Longitudinal Employer-Household Dynamic Program (LEHD); Census Bureau	Connects establishment data from LBD to household data. 4 million establishments for about 20 states.	See LBD and in addition: Employer human capital, workforce indicators,	Yearly, panel, 2003	Establishment	Everyone on payroll	Not publicly available.▲
Economic Census (EC); Census Bureau	Covers 5 million establishments with more than five employees and a sample of the rest.	Employment, Labor costs, Measures of output, Expenses, city identifiers	Years ending in 2 and 7	Establishment, firm	Everyone on payroll	Not publicly available.▲
Company Organization Survey (COS), also known as Report of Organization ; Census Bureau	Surveys 40,000 multi-unit companies with more than 250 employees, and approximately 10,000 smaller multi unit companies on rotating basis.	Establishment operational status, payroll, employment, controlling interests held by other domestic or foreign-owned organizations	Annually since 1974, cross-section, survey coverage and content vary during the census year	Establishment, firm	Everyone on payroll	Not publicly available.▲
Survey of Women/Minority Owned Business Enterprises (SWOBE / SMOBE); Census Bureau	Sample from BITS data frame, Part of EC	Organizational form, Sales and receipts, employees and annual payroll.	1992, 1997, 2002, cross-section	Establishment	Everyone on payroll	Not publicly available.▲
Characteristics of Business owners – both firms and owners (CBO); Census Bureau	Sample from BITS data frame, 78,000-115,000 records for establishments and 117,000-128,000 observations for owners file, Part of EC	Legal form of organization, receipts, sources of capital, employment, whether the business is home based or not.	Yearly, 1982, 1987, 1992; combined with the MOB/WOB in 2002 to form SBO, cross-section	Establishment and individuals	All employees reported on a firm's payroll during specified pay periods	Not publicly available.▲

▲Researchers can apply for access to confidential data <http://www.ces.census.gov/ces.php/rdc> (accessed June 9, 2005)

Name of the dataset; Source	Collection method and coverage	Topic, Main Variables	Periodicity, Dates available, longitudinal links	Unit of observation	How employee is defined	Major limitations
Survey of Business Owners (SBO); Census Bureau	Sample from BITS data frame, 78,000-115,000 records	Legal form of organization, receipts, sources of capital, employment, whether the business is home based or not.	Yearly, cross-section, 2002, for other years see SMOBE /SWOBE	Establishment	All employees on payroll	Not publicly available.
Statistics of Income (SOI); IRS	Stratified probability samples of master file of all tax returns	Tax related issues, Business Receipts, Selected Deductions, Payroll, and Net Income	Yearly, cross-section, 1990-2002	Firm	All employees on payroll	Not publicly available.
Survey of Small Business Finances (SSBF); Federal Reserve Board	Sampling frame is the Dun's Market Identifier (DMI) file. Firms with fewer than 500 employees. About 3,500 businesses.	Firm's use of credit, firm's assets, liabilities, income, revenues, profits, expenses, employment, owners' characteristics	1987, 1993, 1998, cross-section	Firm	Employee on payroll or not, family members on payroll	Only 33% response rate.
National Employer Health Insurance Survey (NEHIS); Centers for Disease Control	National probability sample survey of business establishments, governments, and self-employed individuals with no employees and no other locations. 34,604 completed interviews (70 percent response rate).	Health insurance offerings, employment	1994, cross-section	Establishment	All employee on payroll	Not publicly available. [▲]
Research dataset derived from Martindale-Hubbell Law Directory	Directory of Lawyers and Law firms. Complex algorithms can be used to extracts and match records of lawyers to establishments and firms.	Size of firm, specialization, ratings, full address	1993, 1999, panel	Law firm, law office	Lawyers or supporting personnel affiliated with the firm	Not publicly available

*Researcher can apply for the access to the confidential micro data. For details see <http://www.bls.gov/bls/blsresda.htm> (accessed June 8, 2005).

[▲]Researchers can apply for access to confidential data <http://www.ces.census.gov/ces.php/rdc> (accessed June 9, 2005)

■Researchers can apply for access to confidential data <http://www.cdc.gov/nchs/about/major/nehis/pudf.htm> (accessed June 9, 2005)

Name of the dataset; Source	Collection method and coverage	Topic, Main Variables	Periodicity, Dates available, longitudinal links	Unit of observation	How employee is defined	Major limitations
Kaiser-HRET Employer Health Benefits Surveys ; Kaiser Family Foundation and Research and Educational Trust	Survey of public and private employers, sampled from DMI; about 3,262	Employer health plans coverage, costs, enrollment patterns, health plan choice, and employee costs, employment	Annually since 1999, before that the survey was conducted by KPMG from 1991-1998, and by Health Insurance Association from 1987 till 1991.	Establishment	Not explicitly discussed	Not publicly available. Categorical definition of firm size.
Medical Expenditures Panel Survey (MEPS); Agency for Healthcare Research and Quality	Household Component (HC): Panels of 5 rounds of interviews over 30 months. Insurance component (IC): annual survey of establishments from SSEL sample frame, 27K establishments. Also sample of establishments with workers from prior year HC.	HC: Health status, access to care, income, employment, employment status, eligibility for private and public insurance coverage, health care use and expenses. IC: types of plans provided, number of workers covered, employment: total, by gender, by age over 50, and by earnings.	Annual, 1996-ongoing, HC: panel, IC: cross-section	HC: Household, IC: Establishment	Everyone on payroll, does not include temporary workers	Not publicly available. ■
Duns Market Identifier (DMI); Dunn & Bradstreet (D&B)	Extension of D&B credit database	Info about owners, sales, employment and legal status, full address	Yearly, ongoing, panel can be created using D&B identifiers	Establishment, firm	Includes owners or unpaid family members	Is available for a fee
Kauffman Firm Survey (KFS); Kauffman Foundation	Survey of new businesses from DMI listing sample frame. About 5000 firms.	Owner's characteristics, employment, business organization and benefits, and business finances	Annual, 2005, panel	Owner		Will be publicly available, data have not yet been collected

■ Researchers can apply for access to confidential data <http://www.cdc.gov/nchs/about/major/nehis/pudf.htm> (accessed June 9, 2005)